Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 354 – 363

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Improving Performance of Text Summarization

S.A.Babar[a],Pallavi D.Patil[b*]

[a]*Sanjeevan Engineering & Technology Institute, Panhala*
[b]*Dynaganaga College of Engineering & Research, Narhe, Pune*

Abstract

Today, the tremendous information is available on the internet; it is difficult to get the information fast and most efficiently. There are so many text materials available on the internet, in order to extract the most relevant information from it, we need a good mechanism. Text summarization technique deals with the compression of large document into shorter version of text. Text summarizations choose the most significant part of text and create coherent summaries that state the main purpose of the given document. Extraction based text summarization involves selecting sentences of high relevance (rank) from the document based on word and sentence features and put them together to generate summary. This is modeled using Fuzzy Inference System. The summary of the document is created based upon the level of the importance of the sentences in the document. This paper focuses on the Fuzzy logic Extraction approach for text summarization and the semantic approach of text summarization using Latent Semantic Analysis.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Text summarization; Feature Extraction; Fuzzy logic; fuzzy rule; Latent Semantic Analysis.

1. Introduction

Before going to the Text summarization, first we have to know what a summary is. A summary is a short form of text that is formed from one or more texts that gives important information in the original text[1]. The purpose of automatic text summarization is presenting the source text into a shorter version with semantics[2]. Summary reduces the reading time. There are two types of text summarization methods which are classified into extractive and abstractive summarization[1]. An extractive summarization method is used for selecting important sentences,

* Corresponding author. Tel.: 02342224411;
E-mail address: samrat.babar@seti.edu.in

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.031

paragraphs from the original document. It then concatenates all selected sentences into shorter form. An Abstractive summarization is used to understanding the main concepts in a given document and then expresses those concepts in clear natural language.

There are two different groups of text summarization[2]: Indicative and Informative. Inductive summarization gives the main idea of the text to the user. The length of this type of summarization is around 5 percent of the given text. The informative summarization system gives brief information of the main text .The length of informative summary is around 20 percent of the given text. The automatic summarization means an automatically summarized output is given when an input is applied. Remember that input is well structured document. For this there are initially preprocesses[1] such as Sentence Segmentation, Tokenization, Removing stop words and Word Stemming. Sentence Segmentation is separating document into sentences. Tokenization means separating sentences into words. Removing stop words means removing frequently occurring words such as a, an, the etc. And word stemming means removing suffixes and prefixes. After preprocessing each sentence is represented by attribute of vector of features. For each sentence there are 8 features and each feature has a value between 0 and 1. The 8 features are: Title features, Sentence length, Term weight, Sentence position, Sentence to sentence similarity, Proper noun, thematic word and Numerical data. Our approach is as follows: After extraction of 8 features the result is passed to fuzzifier then to inference engine and finally to defuzzifier. Rules for Inference engine is supplied from Fuzzy rule base. After this each sentence will have score and the sentence is sorted in the decreasing order of the score.

2. Related works

The first Automatic text summarization was created by Luhn in 1958[1] based on term frequency**.** Automatic text summarization system in 1969, has used some standard keyword method as frequency depending weights, cue method, title method and location method are used to assign sentence weights. In 1995,the Trainable Document[2] Summarizer perform sentence extracting task which is based on a number of weighting heuristics.

In 1990s the machine learning techniques in Natural Language Processing used statistical techniques to produce document summaries. They have used a combination of appropriate features and learning algorithms. Hidden Markov models and log-linear models are used to improve extractive summarization**.** Now a day's neural networks are used to generate summary for single documents using extraction. Ladda Suanmali[4] in his work has used sentence weight, a numerical measure assigned to each sentence and then selecting sentences in descending order of their sentence weight for the summary. Recently, neural networks are used to generate summary for single documents using extraction[6].A lot of work has been done in single document and multi document summarization using statistical methods. A lot of researchers are trying to apply this technology to a variety of new and challenging areas, including multilingual summarization and multimedia news broadcast.

3. Motivation

Text Summarization is an active field of research in both the Information Retrieval and Natural Language Processing communities. Text Summarization is increasingly being used in the commercial sector such as Telephone communication industry, data mining of text databases, for web-based information retrieval, in word Processing tools. Many approaches differ on the behaviour of their problem formulations. Automatic text summarization is an important step for information management tasks. It solves the problem of selecting the most important portions of the text. High quality summarization requires sophisticated NLP techniques.

4. Approaches to summarization

Text summarization approach[5] consists of following stages:

1. Preprocessing
2. Feature Extraction

3. Fuzzy logic scoring
4. Sentence selection and Assembly

4.1 Preprocessing:

There are four steps in preprocessing. Segmentation is a process of dividing a given document into sentences. Stop words are removed from the text. Stop words are frequently occurring words such as 'a' an', the' that provides less meaning and contains noise. The Stop words are predefined and stored in an array. Tokenization will separate the input text into separate tokens. Punctuation marks, spaces and word terminators are the word breaking characters. Word Stemming is used to converts every word into its root form by removing its prefix and suffix for comparison with other words.

4.2 Feature Extraction:

The text document is represented by set, D= {S1, S2, - - - , Sk} where, Si signifies a sentence contained in the document D. The document is subjected to feature extraction. The important word and sentence features to be used are decided .This work uses features such as Title word, Sentence length, Sentence position, numerical data, Term weight, sentence similarity, existence of Thematic words and proper Nouns .

1. Title word: A high score is given to the sentence if it contains words occurring in the title as the main content of the document is expressed via the title words. This feature is computed as follows:

$$F1=\frac{Nt}{Ntotal}$$

2. Sentence Length: Eliminate the sentences which are too short such as datelines or author names. For every sentence the normalized length of sentence is calculated as:

$$F2=\frac{No.of the words \in the sentence}{No.of words \int he longest sentnce}$$

3. Sentence Position: The sentences occurring first in the paragraph have highest score. If paragraph has n sentences, the score of each sentence is calculated:

$$F3(S1)=\frac{n}{n};F3(S2)=\frac{4}{5};F3(S3)=\frac{3}{5};F3(S4)=\frac{2}{5};$$ and so on.

4. Numerical data: The sentences having numerical data can imitate important statistics of the document and then selected for summary. Its score is calculated as:

$$F4(Si)=\frac{No.of the Numerical data \in the sentence Si}{Sentnce Length}$$

5. Thematic words: These are domain specific words with maximum possible relativity. The ratio of the number of thematic words that occurs in a sentence over the maximum number of thematic words in a sentence gives the score of each feature as:

$$F5(Si)=\frac{No.of thematic data \in the \in the sentence Si}{Max.no.of thematic words}$$

6. Sentence to Sentence Similarity: Token matching method is used to compute similarity between each sentences S and every other sentences .The matrix [N][N] is formed. N is the total number of sentence in a

document. The diagonal elements of a matrix are set to zero as the sentence should not be compared with itself. The similarity of each sentence pair is calculated as follows:

$$F6 = \frac{\sum[(Si,Sj)]}{MAX[(Si,Sj)]}, \text{ Where i=1 to N and j=1 to N.}$$

7. Term weight: The ratio of summation of term frequencies of all terms in a sentence over the maximum of summation values of all sentences in a document gives the score of term weight feature. It is calculated by the following equation.

$$F7 = \frac{\sum TFi}{MAX \sum TFi}$$

8. Proper Nouns: The important sentence is that sentences which contains maximum number of proper nouns. Its score is given by,

$$F8 = \frac{No.\,of propernouns \in thesentenceSi}{SentencelengthofSi}$$

4.3 Fuzzy Logic Scoring:

Thus each sentence is associated with 8 feature vector. Using all the 8 feature scores, the score for each sentence are derived using fuzzy logic method. Fuzzy rules and triangular membership functions are used in the fuzzy logic method. The fuzzy rules are in the form of IF-THEN .The triangular membership function fuzzifies each score into one of 3 values that is LOW, MEDIUM & HIGH. Then we apply fuzzy rules to determine whether sentence is unimportant, average or important. This is also known as defuzzification.

For example[4] IF F1is High and F2 is Medium and F3 is High and F4 is Medium and F5 is Medium and F6 is Medium and F7 is High and F8 is High THEN sentence is important.

4.4 Sentence Selection:

Based on sentence scores all the sentences are ranked in descending order. Sentences with highest score are extracted as document summary. Finally the sentences in summary are arranged in the order they occur in the original document.

5. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a statistical model[18] of word usage which compares the semantic similarity between pieces of textual information. LSA is designed to improve the effectiveness of information retrieval methods by using "semantic" content of words in a query as opposed to performing direct word matching. LSA avoids the problems of synonymy, in which different words can be used to describe the same semantic concept.

There are three main steps in Latent Semantic Analysis[18]. These steps are as follows:

1. Input Matrix Creation.
2. Singular Value Decomposition.
3. Sentence Selection.

5.1 Input Matrix Creation:

The input document is represented in a matrix form to perform the calculations. A matrix is created which represents the input text. The sentences of the input text are represented by the columns of the matrix and the rows

represent the words in the sentences. The cells of matrix represent the importance of words in sentences. The created matrix is sparse.

The first step of input matrix creation is to create the matrix in the form of terms x sentences. The matrix A with size of m x n is created,where m represents the terms and n represents the sentences, which is M = [M1, M2, Mn]. Each column Mi represents weighted term vector of sentence i of the input document. The terms can be words/phrases that have been seen in the sentences, or they can be preprocessed before the creation of the matrix. The cell represents the frequency of the word in the sentence.

5.2 Singular Value Decomposition:

For modeling the relationship among words/phrases and sentences, singular value decomposition is used. Singular value decomposition is a mathematical method[18] which models the relationships among terms and sentences. It decomposes the input matrix into three other matrices as follows:

$$A=U\sum V^T$$

A: Input matrix (m x n)

U: Words x Extracted Concepts (m x n)

$\sum$: Scaling values, diagonal descending matrix (n x n)

V^T: Sentences x Extracted Concepts (n x n)

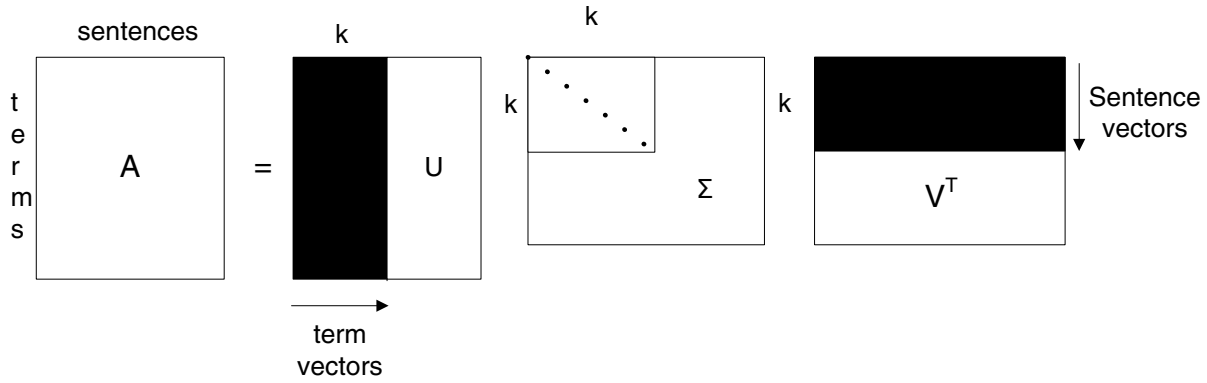

Fig. 1. Singular Value Decomposition[18].

4.3 Sentence Selection:

To select important sentences using the singulav value decomposition results, there are different algorithms proposed. Different algorithms such as,Gong and Liu approach 2001, Steinberger and Jezek's approach 2004 and Murray, Renals and Carletta 2005 are proposed to select sentences using the results of SVD. We have used Gong and Liu, 2001 summarization algorithm which use matrix V^T for sentence selection.

6. Proposed method for Summarization

Traditional extraction methods cannot extract hidden semantic relations between concepts in a text. Therefore, we have used the latent semantic analysis to capture those semantic contents in sentences along with sentence

extraction method to bring the improved summary. Our proposed method can improve the quality of summary with the help of latent semantic analysis and sentence feature extracted fuzzy logic system.

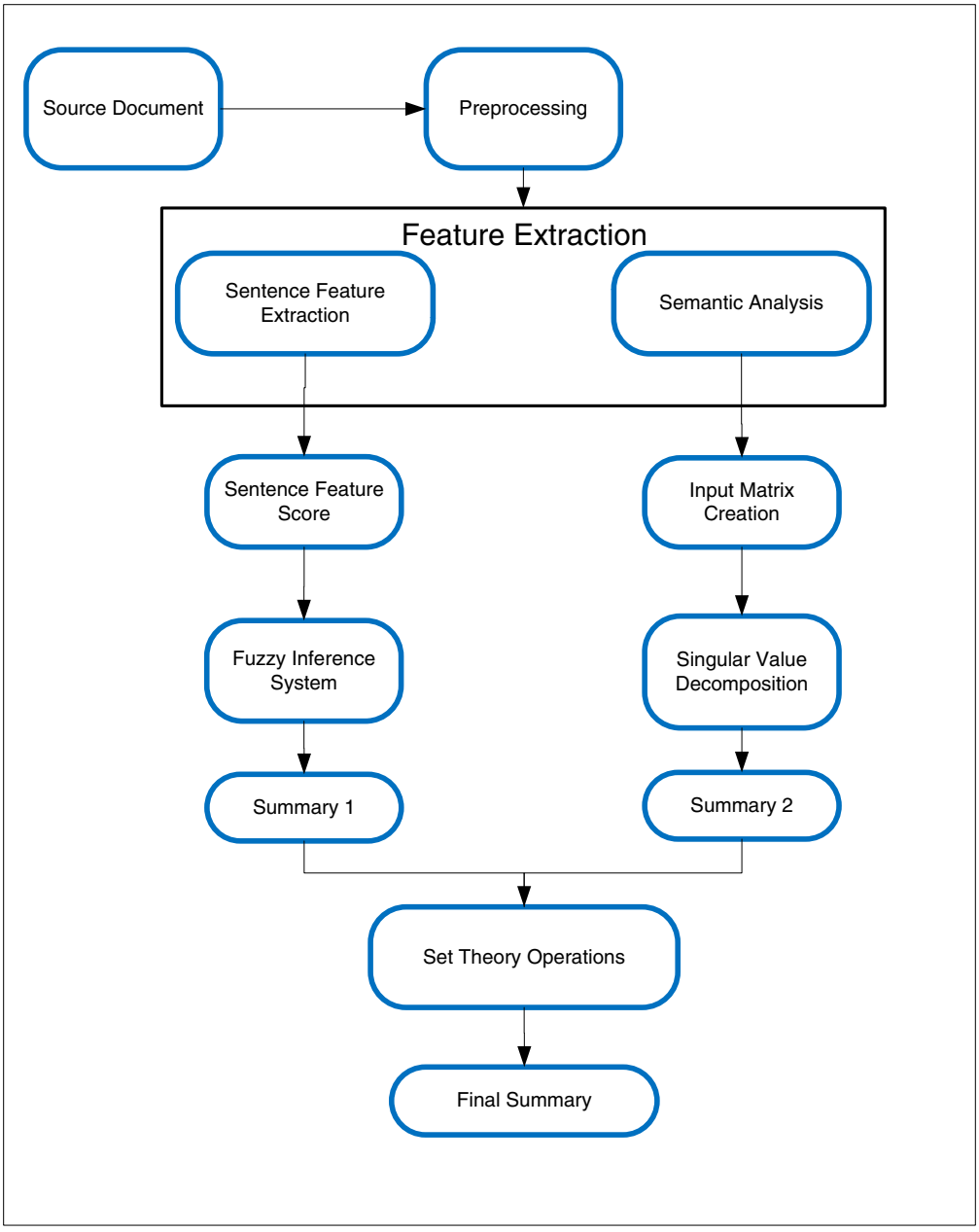

Fig.2. Proposed Architecture.

The system consists of the following main steps:

1. System reads the original source document; 2.For preprocessing step, the system extracts the each sentences of the original documents. It separates the input document into each words. Next, it removes stop words. Word

stemming is the last step for preprocessing; 3. In the sentence features extraction of fuzzy system each sentence is related with vector of eight features that described in above Section, the values are derived from the content of the sentence; in the same way to the semantic system the input matrix of term by documents is created with cell values. 4. Then in the fuzzy system the features are calculated to obtain the sentence score by fuzzy logic method shows in Figure 1; and singular value decomposition is performed on input matrix to obtain the sentence by extracted concept matrix. i.e. importance of the concept & sentences related to that concept. Sentence with highest index value is the most representative sentences of that concept. 5. In fuzzy system, highest score sentences are extracted as document summary based on the compression rate, and in SVD system, the V^T matrix cell values represents the most important sentences extracted. A higher cell value indicates the most related sentence .Thus numbers of sentences are collected into the summary based on compression rate. 6. After getting summary1 and summary2, we intersect both summaries and extract a set of common sentences and a set of uncommon sentences. From uncommon set, we extract the sentences with high sentence scoring. And final set of improved summary is obtained by union of both the sets. We have incorporated the semantic contents of sentences into the sentence feature extraction Fuzzy summarization method to bring the much more improved summary.

6.1 Mathematical Expression:

D = {s1 , s2 , s3 , s4,.....sn}

Where - s is Sentences.

- D is Document.

If (W>T)

Where - W is Weight.

- T is Threshold.

Sum (old) = {s1 , s2};

Sum (new) = {s1 , s3 , s4};

Sum (Final) ={ Sum(old) ^ Sum(new)} U { WP [Sum(old) U Sum(new)]};

Where ..... (W>T)

Sum (Final) = {s1} U {s4} ;

Sum (Final) = {s1, s4};

Thus Summary is improved with our proposed method .Sentences are selected in the summary with the help of sentence scores. Higher scoring (ranked) sentences are added into the summary. Summary from Fuzzy system S1 (Sum (old)) and summary from LSA S2 (Sum (new)) are taken into account and common sentences are kept in one set and other sentences from S1 and S2 are chosen by their sentence scores. Sentences with high score are added into the summary.

7. Experimental Result

Our summary correlates highly with human judgment and has high recall and precision significance test with manual evaluation results. We choose precision, recall as the measurement of our experiment results. We have used ten different data sets.

Table 1. Precision, recall, f-measure Values of fuzzy based Summary.

Datasets	Precision	Recall	f-measure
Dataset(1)	87.8787	41.4285	64.6536
Dataset(2)	92.3076	42.7571	67.2824
Dataset(3)	91.1764	47.6923	69.4343
Dataset(4)	82.5786	33.3333	58.0459
Dataset(5)	95.238	44.4444	69.8412
Dataset(6)	87.8787	46.031	66.9552
Dataset(7)	87.096	45.7527	66.4297
Dataset(8)	84.8484	43.0769	63.9627
Dataset(9)	79.1666	41.3043	60.2355
Dataset(10)	80.9523	30.6363	59.7943
Average	86.91213	41.64568	64.66348

The average precision of fuzzy based summary is 86.91213; average recall is 41.6456 and the average f-measure is 64.6634.The respective graphs of each precision, recall and f-measure are given bellow.

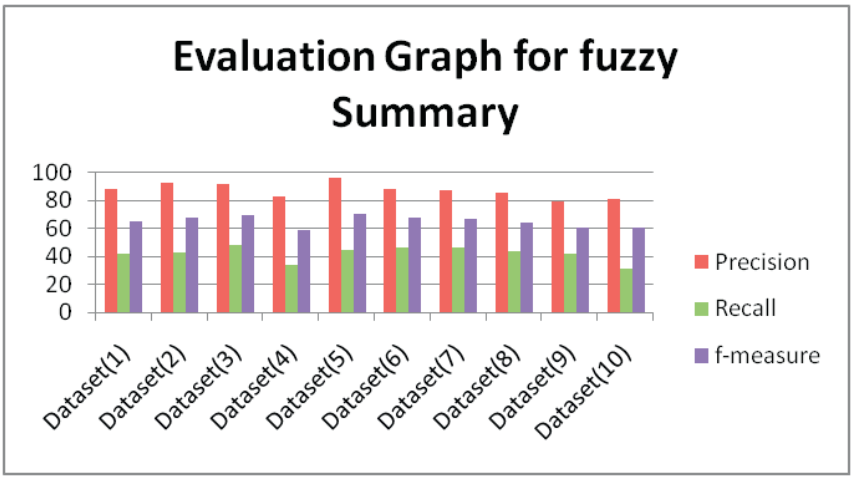

Fig. 3.Evaluation graph of Fuzzy based Summary.

The average precision of proposed summary is 90.77572; average recall is 44.36375and the average f-measure is 67.56974.The respective graphs of each precision, recall and f-measure are given below.

Table 2. Precision, recall, f-measure Values of Proposed Summary

Datasets	Precision	Recall	f-measure
Dataset(1)	90.909	42.857	66.883
Dataset(2)	92.307	42.857	67.582
Dataset(3)	94.117	49.231	71.674
Dataset(4)	86.206	34.722	60.464
Dataset(5)	95.238	44.444	69.841
Dataset(6)	96.9696	50.7936	73.8816
Dataset(7)	90.3225	47.4576	68.8901
Dataset(8)	87.8787	44.6153	66.247
Dataset(9)	83.3333	43.4782	63.4057
Dataset(10)	90.4761	43.1818	66.829
Average	90.77572	44.36375	67.56974

The table 2 represents the precision, recall and F-measure of ten datasets which we have used foe fuzzy based summary. The respective graphs of precision, recall and f-measure are given.

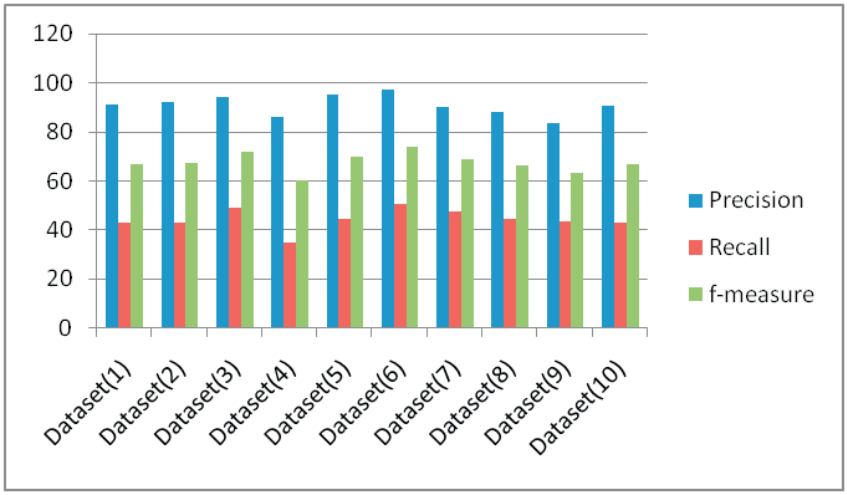

Fig.4.Evaluation graph of Proposed Summary.

The graph shows that the average precision value of proposed summary is 90.77572, recall of 44.36375 and f-measure of 67.56974; while old summary (Fuzzy based summary) summarizer gives the average precision 86.91213, recall of 41.6456and f-measure of 64.6634. Overall precision, recall and f-measure score from our proposed summarizer are better than fuzzy based summary shown in graph.

The result in the graph shows that our proposed summarizers perform better than fuzzy summarizer approach. With precision, recall and f-measure results we have compared the performance of our summarizer with other summarizer.

7. Conclusion

Automatic summarization is a complex task that consists of several sub-tasks. Each of the sub-tasks directly affects the ability to generate high quality summaries. In extraction based summarization the important part of the process is the identification of important relevant sentences of text. Use of fuzzy logic as a summarization sub-task improved the quality of summary by a great amount. The results are clearly visible in the comparison graphs. Our algorithm shows better results as compared to the output produced by two online summarizers.

Thus our proposed method improves the quality of summary by incorporating the latent semantic analysis into the sentence feature extracted fuzzy logic system to extract the semantic relations between concepts in the original text. The focus of this paper is narrow: summarization of documents, but the ideas are more broadly applicable. We need to extend the proposed method for multi document summarization with a large data sets and domain specific data.

References

1 . Saeedeh Gholamrezazadeh ,Mohsen Amini Salehi. *A Comprehensive Survey on Text Summarization Systems*, 978-1-4244-4946-0; 2009 IEEE.

2 . Vishal Gupta, Gurpreet Singh Lehal. *A survey of Text summarization techniques* , Journal of Emerging Technologies in Web Intelligence VOL 2 NO 3 ;August 2010.

3 . Oi Mean Foong, Alan Oxley, Suziah Sulaiman. *Challenges and Trends of Automatic Text Summarization*, International Journal of Information and Telecommunication Technology; Vol.1, Issue 1, 2010.

4 . Archana AB, Sunitha. C . *An Overview on Document Summarization Techniques*, International Journal on Advanced Computer Theory and Engineering (IJACTE) ; ISSN (Print) : 2319, 2526, Volume-1, Issue-2, 2013.

5 . Rafael Ferreira ,Luciano de Souza Cabral ,Rafael Dueire Lins ,Gabriel Pereira e Silva ,Fred Freitas ,George D.C. Cavalcanti ,Luciano Favaro. *Assessing sentence scoring techniques for extractive text summarization*, Expert Systems with Applications 40 (2013); 5755-5764, 2013 Elsevier.

6 . L. Suanmali , N. Salim and M.S. Binwahlan. *Fuzzy Logic Based Method for Improving Text Summarization*, International Journal of Computer Science and Information Security; 2009, Vol. 2, No. 1, pp. 4-10.

7 . Mrs.A.R.Kulkarni, Dr.Mrs.S.S.Apte. *A DOMAIN-SPECIFIC AUTOMATIC TEXT SUMMARIZATION USING FUZZY LOGIC*, International Journal of Computer Engineering and Technology (IJCET); ISSN 0976- 6367(Print), ISSN 0976 - 6375(Online) Volume 4, Issue 4, July-August (2013).

8 . Farshad Kyoomarsi ,Hamid Khosravi ,Esfandiar Eslami ,Pooya Khosravyan Dehkordy. *Optimizing Text Summarization Based on Fuzzy Logic*, Seventh IEEE/ACIS International Conference on Computer and Information Science; 9780-7695-3131-1, 2008.

9 . Ladda Suanmali ,Naomie Salim and Mohammed Salem Binwahla. *Feature-Based Sentence Extraction Using Fuzzy Inference rules*, International Conference on Signal Processing Systems; 978-0-7695-3654-5, 2009 IEEE.

10 . Ladda Suanmali, Naomie Salim , Mohammed Salem Binwahlan. *Fuzzy Genetic Semantic Based Text Summarization*, 2011 Ninth International Conference on Dependable, Autonomic and Secure Computing; 978-0-7695-4612-4, 2011 IEEE.

11 .Ladda Suanmali, Mohammed Salem Binwahlan ,Naomie Salim. *Sentence Features Fusion for Text Summarization Using Fuzzy Logic* ,2009 Ninth International Conference on Hybrid Intelligent Systems ;978-0-7695-3745-0 ,2009 IEEE.

12 . Hsun-Hui Huang ,Yau-Hwang Kuo ,Horng-Chang Yang. *Fuzzy-Rough Set Aided Sentence Extraction Summarization*, Proceedings of the First International Conference on Innovative Computing; Information and Control (ICICIC'06),0-76952616-0/06 ,IEEE.

13 . Feifan Liu and Yang Liu. *Exploring Correlation Between ROUGE and Human Evaluation on Meeting Summaries*, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING; VOL. 18, NO. 1, JANUARY 2010.

14 . ZHANG Pei-ying, LI Cun-he. *Automatic text summarization based on sentences clustering and extraction*, 978-1-4244-4520-2 ; 2009 IEEE.

15 . Udo Hahn, Inderjeet Man. *The Challenges of Automatic Summarization*, 00189162/00;2000 IEEE .

16 . Róbert Móro, Mária Bieliková. *Personalized Text Summarization Based on Important Terms Identification*, 2012 23rd International Workshop on Database and Expert Sytems Applications ;1529-4188, 2012 IEEE .

17 . Rafael Ferreira, Luciano de Souza Cabral, Rafael Dueire Lins, Gabriel Pereira e Silva. *Assessing sentence scoring techniques for extractive text summarization*, Expert Systems with Applications 40 (2013) 5755-5764; 2013 Elsevier Ltd.

18 . Rasha Mohammed Badry, Ahmed Sharaf Eldin ,Doaa Saad Elzanfally. *Text Summarization within the Latent Semantic Analysis Framework: Comparative Study*,International Journal of Computer Applications (0975 – 8887);Volume 81 – No.11, November 2013.